Expanding functional protein sequence spaces using generative adversarial networks

Year: 2021 | Citations: 184 | Authors: Donatas Repecka, Vykintas Jauniškis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rok

---

Abstract

De novo protein design for catalysis of any desired chemical reaction is a long-standing goal in protein engineering because of the broad spectrum of technological, scientific and medical applications. However, mapping protein sequence to protein function is currently neither computationally nor experimentally tangible. Here, we develop ProteinGAN, a self-attention-based variant of the generative adversarial network that is able to ‘learn’ natural protein sequence diversity and enables the generation of functional protein sequences. ProteinGAN learns the evolutionary relationships of protein sequences directly from the complex multidimensional amino-acid sequence space and creates new, highly diverse sequence variants with natural-like physical properties. Using malate dehydrogenase (MDH) as a template enzyme, we show that 24% (13 out of 55 tested) of the ProteinGAN-generated and experimentally tested sequences are soluble and display MDH catalytic activity in the tested conditions in vitro, including a highly mutated variant of 106 amino-acid substitutions. ProteinGAN therefore demonstrates the potential of artificial intelligence to rapidly generate highly diverse functional proteins within the allowed biological constraints of the sequence space. A protein’s three-dimensional structure and properties are defined by its amino-acid sequence, but mapping protein sequence to protein function is a computationally highly intensive task. A new generative adversarial network approach learns from natural protein sequences and generates new, diverse protein sequence variations, which are experimentally tested.